国外平台型媒体虚假信息的技术治理措施及启发

申金霞 张佳迎
（中国传媒大学，北京 100024）

摘 要： 当前，互联网空间治理已经成为全球的重要议题。虚假信息在平台型媒体中进行裂变式传播，强化平台责任、发展技术治理手段成为共识。为借助技术力量对海量以及自动化生产并传播的虚假信息进行检测，近年来，国外的平台型媒体积极采用先进技术打击平台虚假信息，如贴标签增加虚假信息辨识度、通过 AI 实现对虚假信息的自动化检测和标注；通过算法提高权威信息排名或降低虚假信息排名；通过机器学习检测图片、视频和深度伪造，以及识别社交机器人等。文章梳理了目前国外相关的技术治理措施，为我国平台型媒体通过技术手段打击虚假信息提供参考与借鉴。

关键词： 虚假信息；平台型媒体；深度伪造；社交机器人；平台治理　　**中图分类号：** G210　　**文献标识码：** A
文章编号： 1671-0134（2021）08-007-06　　DOI：10.19483/j.cnki.11-4653/n.2021.08.001
本文著录格式： 申金霞，张佳迎 . 国外平台型媒体虚假信息的技术治理措施及启发 [J]. 中国传媒科技，2021（08）：7-12.

导语

平台型媒体（Platisher）的概念最早由乔纳森・格里克（Jonathan Glick）在《平台型媒体的崛起》（Rise of the Platishers）一文中提出，它将平台（Platform）和出版商（Publisher）的优势结合在一起。Digiday 的一位撰稿人将平台型媒体进一步清晰地界定为“既拥有媒体的专业编辑权威性，又拥有面向用户的、开放的平台数字内容实体。”[1] BuzzFeed、Medium、今日头条等网站，以及 Facebook、Twitter、YouTube、TikTok、微信等社交媒体平台都具有平台型媒体的特性，它们将特定的算法技术与专业的编辑运作结合起来，在内容的生产、聚合和分发方面产生巨大的能量。因此，平台型媒体既具有科技平台的开放特质，也具有媒体出版领域的把关属性。[2]

平台型媒体已经成为用户获取信息的主要渠道。与此同时，平台型媒体中的标题党、虚假信息、低俗内容等问题屡禁不止，对网络舆论走向和互联网空间治理产生消极影响，这也是平台型媒体被诟病的焦点所在。虚假信息在平台型媒体中的传播呈现出越来越复杂的样貌，图像、音视频类的虚假信息传播量大，且其留存的副本很难彻底删除。2020 年美国大选期间，一段最初发布在 TikTok 的视频显示：加利福尼亚州的一个选民为民主党候选人乔・拜登填写了七张选票。虽然制作者澄清这些选票只是样票，并且这段视频因违反 TikTok 社区规定被删除，但其副本仍在 Twitter、Facebook 等平台广泛传播，并被作为选举欺诈的证据。[3]

社交机器人（Social Bots）作为一种自动化帐户，能够将信息的传播力进行数量级扩大。近年来，Twitter 社交机器人被某些政治集团利用扩散虚假信息，以影响美国大选等政治事件中的舆论走向。[4] 奥巴马在视频中大骂特朗普、马克・扎克伯格吹嘘自己“完全控制数十亿人的失窃数据”等“深度伪造”（Deepfake）视频引起了广泛关注，深度伪造通过让两个神经网络相互博弈的方式实现自进化，生成内容难以识别真伪。

现实情况倒逼平台承担更多的责任，平台治理也成为全球新闻传播领域的重要议题。目前我国对平台型媒体虚假信息的研究，集中于分析某个具体事件中虚假信息的传播机制，或是从政府治理、政策法规制定以及用户媒介素养入手进行讨论，技术维度的平台治理有待深化。因此，梳理国外主要平台型媒体的技术治理措施，对于我国的互联网内容治理具有重要的借鉴意义。

1. 国外平台型媒体虚假信息的技术治理措施

为了打击虚假信息，提高平台信息的整体可信度，国外平台型媒体的主要技术治理措施包括标签（Label）、排名（Ranking）、人工智能检测等措施。

1.1　标签＋ AI 自动化检测

标签是指平台将额外的信息附加在用户生成内容中的一种手段，这些信息包括事实核查结果、内容警告或更多的上下文信息，用户可以根据标签的提示判断信息真伪，最常用的是“可信度标签”（Credibility Labels）和上下文标签（Contextual Labels）。[5] 可信度标签提供了关于信息真伪的明确信息，Facebook 自 2020 年 3 月开始，便使用这类标签，标记了超过 1.8 亿个有关美国大选的虚假信息的帖子，[6] 平台用灰色背景覆盖原帖内容，并添加“False Information”的警告标签，直接告诉用户该帖子经事实核查后确认为虚假信息，并附上揭露该帖子的详细内容的链接。上下文标签只提供给用户更多的

基金项目： 本文系中国广播电视社会组织联合会项目“媒体融合背景下虚假新闻舆情传播及治理研究”（项目编号：2020ZGLH007）的研究成果。

相关信息，不对内容可信度做出任何明确的声明或判断，目的是鼓励用户在了解更多背景信息后做出自己的判断。TikTok 在 2020 年 12 月推出新标签 #covidvaccine（#疫苗），以检测和标记所有与 COVID-19 疫苗相关的视频，并在视频中附加横幅“Learn more about COVID-19 vaccines”（“了解有关 COVID-19 疫苗的更多信息”），将用户引向可验证的、权威的信息来源。[7] 不同平台对标签的术语、位置、视觉和交互设计有所不同，Twitter 将虚假信息标签分为误导性信息（Misleading information）、有争议的声明（Disputed claims）以及未经证实的声明（Unverified claims），并针对不同类型信息采取删除、警告等不同的措施。

平台型媒体通过第三方事实核查机构检测与自动化技术相结合的方式来实现对海量信息的检测和标注。以 Facebook 为例，该平台于 2016 年 12 月启动了事实核查计划，聘请第三方组织对 Facebook 内容的准确性进行评估和审查，截至 2020 年 10 月，Facebook 已经与 80 个事实核查组织建立合作，涵盖全球 60 多种语言。[8] 用户对 Facebook 中的某个帖子表达质疑或进行举报等行为是第三方事实核查组织需要介入的信号。经过事实核查人员评估后确定为虚假信息的内容，会被平台贴上“虚假信息”的标签。Facebook 将事实核查人员审查过的以往文章作为素材，通过机器学习来扩大信息的检测范围，弥补人工审查在数量、速度上的局限性。2020 年 4 月，Facebook 根据事实核查组织检测的约 7500 篇文章，使用人工智能给大约 5000 万条与 COVID-19 相关的内容贴上了警告标签。[9]Twitter 团队也声称正在使用和改进系统，以快速检测和标记与 COVID-19 相关的内容，并确保标签的使用不会扩大虚假信息的传播量。Instagram 则使用图像匹配技术来查找相似的内容并添加标签。此外，Facebook 还通过使用一种专门用于检测接近的重复项算法模型 SimSearchNet 来识别图像类虚假信息的副本，如截屏图片。通过 SimSearchNet 的精准匹配，可以实现对虚假信息的副本进行识别和标记。[10]

有研究表明，添加标签在抑制平台虚假信息传播方面有积极作用。加利福尼亚大学 2019 年的一项研究表明，[11] 通过对虚假信息进行标记能够对用户的分享意图产生影响，进而减少虚假信息的扩散。上下文标签向用户提供了背景介绍、针对错误的解释以及相关权威内容等细节信息，有助于用户纠正自己固有认知与事实真相的不一致性，并促进更持久的信念改变。[12] 此外，添加标签以减少虚假信息传播必须建立在大规模覆盖的基础上，以 Facebook、Twitter 为代表的平台型媒体也正通过人工智能等技术手段，训练系统自动对海量信息进行检测和标记，以克服人工核查的局限性。

1.2 提高优先级与降权

对平台型媒体来说，依靠平台自身的技术架构和算法机制实现内容的聚合与推送，是其区别于传统新闻网站线性呈现新闻的重要特征之一。为了实现用户内容的个性化定制、避免信息过载，平台型媒体采用推荐算法对信息进行过滤，按照平台所制定的规则对信息进行评估、排序和推送，这是一个以算法为核心的自动化决策过程，包括排序、分类、关联和过滤四个步骤（见表 1）。不同的平台型媒体在赋予算法要素的权重上有所不同：NewsFeed 作为 Facebook 的核心新闻推送项目，其算法机制从最初的边际排名算法（EdgeRank Algorithm），到加入用户关系、用户偏好、最近联系人等要素，再到协同过滤机制，可以看出，Facebook 的内容推荐机制是基于用户个人的社交关系和互动偏好。相比之下，在 Google 的推荐逻辑中，内容的主题分类和优质程度是非常重要的指标。而 TikTok 的推荐机制更多是基于用户个人的偏好和活动历史，包括用户的点赞、分享、关注等指标。[13] 尽管不同平台的推荐机制中要素赋权有所差异，但有一点是共通的，即点击越多、推荐越多。

表 1 算法决策的四个步骤 [14]

步骤	内涵
排序（Prioritization）	或称优先级、排名，目的在于突出强调某些事物，而弱化其他事物
分类（Classification）	通过检验事物的部分特征，将其划分至一个特定组别中
关联（Association）	标注不同事物之间的联系
过滤（Filtering）	根据不同的标准包含或排除特定信息

YouTube、Facebook 和 Twitter 等平台型媒体的算法推荐系统，要为虚假信息、仇恨言论、阴谋论和其他有害内容的传播承担很大责任。虚假信息常与具有较高关注度的社会重大事件或政治议题相关联，并具有情绪的煽动性，这类信息更容易引起用户的点击关注。一旦进入推荐系统中，这类信息就会在用户社交关系、附近推荐、流行度、相关主题等要素的作用下被推荐给更多的用户，循环往复形成病毒式传播，而这一切都是基于算法的自动化决策。

解决这一问题的办法就是减少虚假信息进入推荐系统的机会，即降低信息的排名或降权（Downranking），以减少虚假信息在平台中的可见度。在世界卫生组织宣布 COVID-19 成为全球突发公共卫生事件之后，平台型媒体积极采取措施减少虚假信息和有害内容的传播。2020 年 8 月，Facebook 公开了其有关内容推荐的运作方式，标题党、欺骗性的信息及不实或误导性信息等内容会被降权，经常分享虚假信息的用户和群组也会被减少推荐。为了应对 COVID-19 期间的信息疫情，Facebook 会减少被第三方事实核查机构和世界卫生组织判定的虚假信息，包括“夸大或耸人听闻的健康声明，以及那些以健康声明为幌子进行产品或服务销售的人”，同时提高权威信息的排名。[15]Google 也有类似的做法，当搜索的排名算

法识别出信息中包含虚假信息时，会降低该内容的排名，同时提高官方信息的优先级。Instagram 会将其从“探索”和主题标签页面中删除，并降低其在动态和故事中的可见性来减少分发。通过降权的方式，包含虚假信息的内容，在信息流中的排名被降低甚至是无排名，这些信息进入推荐系统的机会大大减少，从而抑制了虚假信息的阅读和传播。

1.3 通过机器学习检测图片、视频及深度伪造

人类进入视觉传播时代，用户每天接收大量的图像、视频等视觉化文本。YouTube 作为一个用户生产的视频类社区，在用户活跃度方面是全球仅次于 Facebook 的平台；[16]TikTok 在全球拥有 6.89 亿月活跃用户，并在 2020 年成为 iPhone 用户下载量第二大的应用程序。[17] 随着用户行为习惯的改变，虚假信息也经历了由“单模态”到“多模态”的发展。路透社用户生成内容的全球负责人黑兹尔•贝克表示：“2019 年几乎所有的全球重大新闻事件中，都在社交媒体上识别出了误导性的视频和图片。”[18]

在此背景下，由人工驱动的事实核查具有很大的局限性，人工难以判别真伪，且在速度上远远滞后于虚假信息的传播数量和速度。

Facebook 在 2019 年将事实核查拓展到对图片和视频的核查，他们将图片视频类的虚假新闻分为三种：篡改（Manipulated or fabricated）、断章取义（Out of context）和包含虚假文本或音频声明（Text or audio claim）。断章取义意味着图像或视频的使用脱离了上下文语境，使得它的真实性发生扭曲。事实核查员可以利用他们的新闻专业知识和情境理解能力识别这一类型的虚假信息。较为复杂的是第一种和第三种类型，必须采用技术手段才能实现对海量信息的检测。

对包含虚假文本或音频的照片来说，第一步是使用光学字符识别技术（Optical Character Recognition，简称 OCR）或音频转录工具提取文本，再通过自然语言处理方法，将提取的文本与事实核查员检测确认的虚假信息进行匹配，查看是否有文本重复内容。Facebook 建构了一个名为 Rosetta 的大型机器学习系统，每天实时从超过 10 亿的 Facebook 和 Instagram 公共图像和视频帧中提取文本，并将其输入到经过训练的文本识别模型中，以帮助机器理解上下文的文字和图片的构成，[19] 这项技术能够帮助 Facebook 实现对其平台上多变且海量的图像进行检测。

经过篡改的图片和视频是最难以识别的，深度伪造也是其中的一种。深度伪造在“生成对抗网络”的基础上，通过让“生成器”和“鉴别器”两个神经网络相互博弈的方式进行学习，在相互对抗、不断调整参数中实现自进化，最终目的是鉴别器无法判断生成器的输出结果是否真实，因此生成的深度伪造图像、视频常常能够以假乱真。为了更好地检测深度伪造，Facebook 与密歇根州立大学的研究团队建立合作，通过逆向工程（Reverse Engineering）来识别和跟踪深度伪造信息。他们要求系统首先假设图像是深度伪造的，然后通过逆向工程算法追踪定位创建它的 AI 软件。首先通过指纹评估网络（Fingerprint Estimation Network）运行深度伪造图像，评估 AI 软件留下的指纹信息，算法知道指纹之后，可以与算法输出端深度伪造视频或图像指纹进行匹配。通过这种方式，可以定位追踪到具体是哪一款软件生成了该深度伪造产品。该系统在关键基准测试中的准确率为 70%，优于之前的任何一种检测方式。[20]

1.4 通过机器学习检测社交机器人

社交机器人（Social Bots）是一种能够自动生成内容的算法程序，在社交网络中能够模仿、影响人的行为，并与用户进行互动。社交机器人最初用于聚合内容、自动回复信息，但随着技术的滥用，被恶意设计，用于传播虚假信息、发送垃圾邮件甚至只是噪音以误导和操纵舆论。[21] 有研究指出，社交机器人占 Twitter 活跃用户总数的 9% 至 15%；[22]Instagram 中的机器人账户多达 9500 万个，占用户总数的 9.5%。[23] 并且和真人相比，社交机器人更乐于分享，有研究分析了 120 万条带有超链接的推文，发现约有 66% 的热门新闻媒体网站的链接是由机器人账户自动发布的。[24]

社交机器人的危害在于它不加核查地发送内容，其中不乏有各类虚假信息、仇恨言论，成千上万的社交机器人一起运作，推动了平台内虚假信息的指数级增长。近年来，社交机器人的信息放大器功能被恶意利用，在线生态系统不断受到恶意自动化账户的威胁。社交机器人被认为影响了西方国家重大政治选举的在线讨论，包括 2016 年美国总统大选和英国脱欧公投。[25] 有研究表示，2020 年 1 月以来，讨论有关 COVID-19 信息的 Twitter 账户中，机器人账户高达 45%，它们捏造了超过 100 种有关病毒的虚假叙述，包括一系列有关医院使用人体模型充当患者以夸大病毒严重性程度的阴谋论，以及新冠病毒的传播与 5G 无线发射塔有关等。[26] 因此，对平台来说，识别、删除这些带有恶意的社交机器人十分重要。

检测社交机器人的技术主要有基于图（Graph-Based）、基于行为特征（Featured-Based）以及众包（Crowdsourcing）三种类型，众包依靠专业人员的人工识别实现，前两种则依靠机器学习实现。[27] 基于图的社交机器人检测技术主要是依据账户的社交网络图来理解和分析平台上用户之间的关系。真实的用户会有大量的关注、转发和双向互动行为，因此呈现出的社交网络图结构会与机器人账户有很大不同。[28] 基于行为特征的检测技术是从多个角度对账户的行为模式特征进行分析。大西洋理事会的数字取证研究实验室（DFRLab）提出，可以通过账户活跃度、匿名性以及发布行为特征等 12 个信号识别政治类社交机器人，行为特征包括原创内容少、

转发量大；不对转发内容进行评论；推文中含有多种语言等。[29]Botometer 是一种典型的基于行为特征来识别社交机器人的工具，由南加州大学和印第安纳大学的研究人员开发，它会读取账户的 1000 多种不同的特征，并对其特征分配 0 到 1 之间的分数，分数越高，是社交机器人的可能性越大。

面对数量庞大的社交机器人和虚假信息，Twitter 也在不断改进自动化检测技术。2019 年，Twitter 收购了一家致力于机器学习和检测虚假信息的 Fabula AI 公司。该公司擅长于几何深度学习（Geometric Deep Learning），可以将机器学习应用于网络结构数据，实现对大且复杂的关系和交互数据集的描述，[30] 这将帮助 Twitter 更好地改善平台信息环境。此外，Twitter 还要求用户完成简单的 reCAPTCHA 过程或密码重置请求，其中文全称是“全自动区分计算机和人类的图灵测试”，要求用户在 OCR 软件无法识别的文字扫描图中进行操作，以识别社交媒体机器人。有团队研究了 2016 年与 2020 年社交机器人对美国大选的影响，团队负责人表示，相比于 2016 年，目前社交机器人的数量已经急剧减少，可能是像 Twitter 这样的平台在检测和删除社交机器人方面取得了成效。[31]

2. 对我国平台型媒体治理虚假信息的启发

2020 年，新冠肺炎疫情大暴发与信息疫情相互交织，加之美国大选引发的虚假信息，网络空间治理更加迫在眉睫。通过强化平台责任、发展技术治理手段、完善法律法规等措施，促进多元协同治理来打击虚假信息、限制暴恐言论已经成为全球共识。对我国而言，互联网空间治理一直是重要话题。互联网的应用和普及引发了新的信息传播环境，内部面临着由社会各阶层矛盾和社会认同引发的负面舆论，外部面临着国际网络话语权的争夺、网络空间安全以及网络意识形态的较量，[32] 在此背景下，中国一直在探索互联网空间治理的路径。习近平总书记在党的十九大报告中提出的“加强互联网内容建设，建立网络综合治理体系，营造清朗的网络空间”，再次强调了互联网空间治理的重要性。随着 5G 和智能时代的到来，平台型媒体将拥有更强大的计算能力和资源，更有条件和义务发挥“在线看门人”的功能。[33] 国外平台型媒体的治理措施，具有一定的启发和借鉴意义。

第一，不断优化、调整算法，针对不同内容采取“提高优先级”或“降权”。虚假信息、标题党、色情低俗内容泛滥等问题是平台型媒体被诟病的焦点，平台也针对这些问题进行了算法调试。有学者指出，今日头条的推荐算法从第一版运行至今，经过了四次大的调整和升级，体现在算法加人工的半自动形式辨别内容质量、优化算法以实现更精准的用户画像以及优化推荐算法的兴趣探索能力三个方面。[34] 除此之外，还可以借助一定的自然语言处理技术，识别可能含有虚假信息的文章，通过提高权威优质来源信息的优先级以及降低劣质内容排名的方式，减少虚假信息的可见度。

第二，借鉴国外检测图像、视频虚假信息的方法，并积极与第三方技术机构、高校等建立合作，探索识别深度伪造的方法。抖音、快手等短视频平台拥有大量用户，以图像、视频的形式传播的虚假信息难以识别，需要借助技术的力量才能实现批量化事实核查，而国内平台在这方面还较为薄弱。2019 年 8 月，一款名为“Zao”的应用程序引发关注，通过简单的操作便可实现“换脸”，在丰富娱乐资讯的同时，也带来了深度造假、侵权等方面的隐患，虽然这一程序后来被禁用，但因网络操作的隐蔽性和低成本，在规范深度伪造技术使用的同时，也要提升对图像、视频等深度伪造技术的打击能力。

第三，平台型媒体与事实核查机构深度合作。首先目前，独立的事实核查平台或辟谣平台信息受众少，缺乏影响力，平台型媒体可与这些平台合作，使用“标签”提高辟谣效果。微博的 @ 微博辟谣和微信的“微信辟谣助手”都通过定期推送的形式向用户发送辟谣信息，但必须是在用户订阅、点击的情况下才能达到辟谣效果。Facebook 等平台常用的“贴标签”的形式，可以放大传播效果，将事实核查结果以标签形式附在含有虚假信息的帖子上，可以帮助用户快速判断信息真伪，一目了然。其次，大力推进自动化事实核查技术的研发与实践。目前，在打击虚假信息方面，微博、微信等平台都采用了一定的技术手段，如微博辟谣是利用技术手段对转发量较高的信息进行监控，发现疑似的虚假信息后，采用深度搜索、求证专家等方式进行查证。微信则主要通过建立技术拦截的方式，经由人民网、果壳网、丁香园等合作者认定为含有虚假信息的内容，在传播过程中会被拦截。[35] 但在与独立的事实核查机构合作、深入研发自动化事实核查技术方面，我国的平台型媒体还应向国外的平台型媒体学习，可以通过专项合作、资金扶持的方法，推进自动化事实核查技术的发展，从而提升打击假新闻的效率。

结语

智媒时代，虚假信息的生产与治理将是一场以技术为核心的“攻防战”。[36] 技术既可以助推虚假信息，也可以阻止虚假信息。深度伪造通过机器学习制造扭曲现实的视频，难以视频辨别真伪；带有恶意的社交机器人遍布不同平台，扩散虚假信息、误导舆论走向；虚假信息在平台型媒体中依靠算法推荐机制裂变式传播，实现指数级增长。这意味着平台型媒体必须通过技术手段，识别海量信息中的虚假信息，以及自动化生产并传播的虚假信息，在检测基础上，通过标记、拦截和删除的方式，遏制虚假信息的传播。但同时技术具有局限性，自动化的虚假信息检测技术也需要借助人的新闻专业性和信息素养进行运作，促进算法技术与政府、平台、新闻媒体、高校、科研机构、公益组织和用户等多元主体的共同参与，

是网络空间治理的必然趋势。媒

参考文献

[1] Jerome Sun. 平台型媒体：来自硅谷的另一种媒体融合[EB/OL]. https：//medium.com/@jeromesun_66925/%E5%B9%B3%E5%8F%B0%E5%9E%8B%E5%AA%92%E4%BD%93-%E6%9D%A5%E8%87%AA%E7%A1%85%E8%B0%B7%E7%9A%84%E5%8F%A6%E4%B8%80%E7%A7%8D%E5%AA%92%E4%BD%93%E8%9E%8D%E5%90%88-f8af5979e6ce.

[2] 焦洁 . 平台型媒体：一种新型的融媒体 [J]. 西部学刊，2015（1）：20.

[3] Mikael Thalen. Eric Trump keeps falling for fake ballot hoaxes[EB/OL]. https：//www.dailydot.com/debug/eric-trump-keeps-falling-for-fake-ballot-hoaxes/.

[4] Caldarelli，G.，De Nicola，R.，Del Vigna，F. et al. The role of bot squads in the political propaganda on Twitter[J]. Communications Physics，2020（1）.

[5] Emily Saltz，Claire Leibowicz. Fact-Checks，Info Hubs，and Shadow-Bans： A Landscape Review of Misinformation Interventions[EB/OL]. https：//www.partnershiponai.org/intervention-inventory/.

[6] Rachel Kraus. Facebook labeled 180 million posts as "false" since March. Election misinformation spread anyway[EB/OL]. https：//sea.mashable.com/tech/13294/facebook-labeled-180-million-posts-as-false-since-march-election-misinformation-spread-anyway.

[7] TikTok. Taking action against COVID-19 vaccine misinformation[EB/OL].https：//newsroom.tiktok.com/en-gb/taking-action-against-covid-19-vaccine-misinformation.

[8] Facebook for Government，Politics and Advocacy. Understanding Facebook's Fact-Checking Program[EB/OL]. https：//www.facebook.com/gpa/blog/misinformation-resources.

[9] Roshan Sumbaly，Mahalia Miller，Hardik Shah，et al. Using AI to detect COVID-19 misinformation and exploitative content[EB/OL]. https：//ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/.

[10] Roshan Sumbaly，Mahalia Miller，Hardik Shah，et al. Using AI to detect COVID-19 misinformation and exploitative content[EB/OL]. https：//ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/.

[11] Andrew Hutchinson. New Study Finds that Flagging False Reports on Facebook May Indeed Reduce their Distribution[EB/OL].https：//www.socialmediatoday.com/news/new-study-finds-that-flagging-false-reports-on-facebook-may-indeed-reduce-t/559968/.

[12] Ecker U，O'Reilly Z，Reid J S，et al. The Effectiveness of Short-Format Refutational Fact-Checks[J]. British Journal of Psychology，2020（1）：36.

[13] Shannon Mullery. How the TikTok Algorithm Works in 2021[EB/OL]. https：//tinuiti.com/blog/paid-social/tiktok-algorithm/.

[14] 方师师 . 算法机制背后的新闻价值观——围绕"Facebook 偏见门"事件的研究 [J]. 新闻记者，2016（09）：41.

[15] Guy Rosen. An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19[EB/OL]. https：//about.fb.com/news/2020/04/covid-19-misinfo-update/.

[16] Statista.Most popular social networks worldwide as of January 2021，ranked by number of active users[EB/OL]. https：//www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[17] Shannon Mullery. How the TikTok Algorithm Works in 2021[EB/OL]. https：//tinuiti.com/blog/paid-social/tiktok-algorithm/.

[18] Hazel Baker. Introducing the Reuters guide to Manipulated media，in association with the Facebook Journalism Project[EB/OL]. https：//www.reuters.com/article/rpb-hazeldeepfakesblog/introducing-the-reuters-guide-to-manipulated-media-in-association-with-the-facebook-journalism-project-idUSKBN1YY14C.

[19] Viswanath Sivakumar，Albert Gordo，Manohar Paluri. Rosetta： Understanding text in images and videos with machine learning[EB/OL]. https：//engineering.fb.com/2018/09/11/ai-research/rosetta-understanding-text-in-images-and-videos-with-machine-learning/.

[20] Jeremy Kahn. Facebook says it's made a big leap forward in detecting deepfakes[EB/OL]. https：//fortune.com/2021/06/16/facebook-detecting-deepfakes-research-michigan-state/.

[21] Ferrara E，Varol O，Davis C，et al. The Rise of Social Bots[J]. Communications of the Acm，2014（7）：96.

[22] Varol O，Fe Rrara E，Davis C A，et al. Online Human-Bot Interactions： Detection，Estimation，and Characterization[J]. 2017（1）：280.

[23] MarketingDive. Instagram may have 95M bot accounts，The Information reports[EB/OL]. https：//www.marketingdive.com/news/instagram-may-have-95m-bot-accounts-the-

information-reports/528141/.

[24] Stefan Wojcik. 5 things to know about bots on Twitter[EB/OL]. https：//www.pewresearch.org/fact-tank/2018/04/09/5-things-to-know-about-bots-on-twitter/.

[25] Bovet，A.，Makse，H.A. Influence of fake news in Twitter during the 2016 US presidential election[J]. Nat Commun，2019（7）：1.

[26] Bobby Allyn. Researchers： Nearly Half Of Accounts Tweeting About Coronavirus Are Likely Bots[EB/OL]. https：//www.npr.org/sections/coronavirus-live-updates/2020/05/20/859814085/researchers-nearly-half-of-accounts-tweeting-about-coronavirus-are-likely-bots.

[27] Ferrara E，Varol O，Davis C，et al. The Rise of Social Bots[J]. Communications of the Acm，2014（7）：100.

[28] 李阳阳，曹银浩，杨英光，金昊，杨阳朝，石珺，李志鹏 . 社交网络机器账号检测综述 [J]. 中国电子科学研究院学报，2021（3）：214.

[29] DFRLab. #BotSpot： Twelve Ways to Spot a Bot[EB/OL]. https：//medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c.

[30] Parag Agrawal.Twitter acquires Fabula AI to strengthen its machine learning expertise[EB/OL]. https：//blog.twitter.com/en_us/topics/company/2019/Twitter-acquires-Fabula-AI.

[31] Giorgia Guglielmi. The next-generation bots interfering with the US election[EB/OL]. https：//www.nature.com/articles/d41586-020-03034-5#ref-CR2.

[32] 梅松 . 国家总体安全观视域下的互联网信息治理研究 [J]. 社会治理法治前沿年刊，2016（0）：101.

[33] 易前良 . 网络平台在内容治理中的“在线看门人”角色 [J]. 青年记者，2020（7）：24.

[34] 喻国明，杜楠楠 . 智能型算法分发的价值迭代：“边界调适”与合法性的提升——以“今日头条”的四次升级迭代为例 [J]. 新闻记者，2019（11）：19-20.

[35] 卢尚青 . 社交媒体谣言的传播机制研究 [D]. 济南：山东大学，2016：38，40.

[36] 张超 . 社交平台假新闻的算法治理：逻辑、局限与协同治理模式 [J]. 新闻界，2019（11）：28.

作者简介：申金霞（1975-），女，河南民权，新闻学博士，副研究员，中国传媒大学互联网信息研究院，研究方向：舆情治理、互联网内容生产与传播；张佳迎（1997-），女，河南南阳，中国传媒大学互联网信息研究院硕士研究生。

（责任编辑：陈旭管）